

Learning New User Formulations in Automatic Directory Assistance.

*Original*

Learning New User Formulations in Automatic Directory Assistance / C., Popovici; M., Andorno; Laface, Pietro; L., Fissore; M., Nigra; C., Vair. - (2002), pp. 1448-1451. (Intervento presentato al convegno IEEE International Conference on Acoustics, Speech and Signal Processing nel May).

*Availability:*

This version is available at: 11583/1413113 since:

*Publisher:*

IEEE

*Published*

DOI:

*Terms of use:*

openAccess

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

# Learning New User Formulations in Automatic Directory Assistance

*C. Popovici, M. Andorno, P. Laface\*, L. Fissore, M. Nigra, C. Vair<sup>^</sup>*

\* Politecnico di Torino, Italy  
{andorno,laface}@polito.it

<sup>^</sup> Loquendo, Torino, Italy  
(Cosmin.Popovici, Luciano.Fissore, Mario.Nigra, Claudio.Vair}@loquendo.com

## Abstract

Telecom Italia has deployed since the beginning of year 2001 a nationwide automatic Directory Assistance (DA) system that routinely serves customers asking for residential and business listings.

DA for business listings is a challenging task: one of its main problems is that customers formulate their requests for the same listing with a great variability.

Since it is difficult to reliably predict a priori the user formulations, in this paper we propose a procedure for detecting, from field data, user formulations that were not foreseen by the designers. These formulations can be added, as variants, to the denominations already included in the system to reduce its failures.

We show that using a large database associating phonetic transcriptions of user utterances with the phone number provided by the automatic service, a completely unsupervised approach detects most of the old formulations. Furthermore, our procedure is able to filter a huge amount of calls routed to the operators, and to detect a limited number of phonetic strings that are good candidate to be included as new formulation variants in the system vocabulary.

## 1. Introduction

DA is the most used service that Telecom operators offer to their customers. This service is expensive because it relies on a multitude of human operators. A strategy for reducing these costs is to reduce the work time of the operators providing information collected by an Automatic Speech Recognizer, another strategy is to rely as much as possible to complete automation.

Several ASR technology providers have worked on problems related to DA, but most of the work has been concentrated on locality and person name recognition [3], [5], for example. This is not surprising because it is difficult to reliably predict a priori the user formulations for business listings. However, since the vast majority of the calls to the DA services are related to business listings, to reduce the service costs, most of the efforts should be devoted to this category of calls.

The second strategy has been selected by Telecom Italia, that has deployed since the beginning of year 2001 a nationwide automatic DA system, jointly developed with Loquendo (formerly CSELT), that routinely serves customers asking for residential and business listings. Whenever the automatic system is unable to terminate the transaction with the customer, the call is routed to a human operator. A description of the system, related to the management of the residential calls, has been presented in [1].

The analysis of the traffic has shown that about 80% of the DA customer accesses are related to business listings, it is important, thus, to improve the percentage of success of the automatic system for this class of calls.

The Loquendo approach to DA for business listing is based on large vocabulary isolated word recognition technology, where the sequence of words of a business listing is concatenated and transcribed as a single word, with possible silences in between. Since the content of the original records in the database does not, typically, match the linguistic expressions used by the callers, a complex processing step is needed for deriving a set of possible formulation variants (FVs) from each original records in the listing book.

A large percentage of users expressions, however, still remain uncovered by the FV database. Thus, in this paper we propose a procedure for detecting, from field data, user formulations that were not foreseen by the designers. These formulations can be added, as variants, to the denominations already included in the system to reduce its failures. In particular we are interested in detecting new formulations for frequently requested listings, for example nicknames of hospitals or other public services, or user requests for the phone number of a popular TV talk show, that of course does not appear in the directory listings.

Our approach is based on partitioning the field data into phonetically similar clusters from which new user formulations can be derived.

The paper is organized as follows: Section 2 gives a short overview of the Loquendo DA system. Section 3 details the generation of formulation variants and the evaluation of their coverage. Sections 4 and 5 present our approach for learning new formulations from the field data, and our conclusions are given in Section 6.

## 2. Loquendo DA system overview

The initial prompt of the system allows the customer to select between an automatic or a more expensive human operated service. The second prompt is for residential or business listings, then the city name is collected and finally the desired denomination. For business requests, the address is possibly asked to the user whenever the system is not confident about the recognized denomination, if there are ambiguities in the database, or whenever it cannot find the information in the list of the most frequently requested listings. The search for a business denomination is carried out first on the list of the most important or most frequently requested listings (Top-List) that can include up to 25K entries. If the search fails the search continues on the list of the denominations associated to the address provided by the user.

The basic technology for this DA application is isolated word recognition, carried out in two steps.

The first step decodes the user utterance by means of a Hybrid HMM-NN model, where the emission probabilities of the HMM states are estimated by a Multi Layer Perceptron. This step generates also the best phonetic string by means of a phone-looped model, and its score.

The second step, based on Continuous Density HMMs, decodes the same utterance with the vocabulary of the N-best hypotheses produced by the first step. The added value of the second step is twofold since the combination of the hypothesis scores of the two steps, not only increases the recognition accuracy, but allows also the computation of a reliability score for the best hypothesis. The dialog manager module uses the reliability and phonetic scores to reject unreliable hypotheses, or to reduce unnecessary request turns.

### 3. Generation of formulation variants

The Italian telephone book listings include more than 25.000.000 records, about 3.500.000 of which are business listings.

Pref: 011      Tel: 5175296 City: TORINO    Prov: TORINO Address: 63, C. VITTORIO EMANUELE II <b>Den: BAR RISTORANTE LA FORCHETTA D'ORO</b> <b>DI MARIO ROSSI</b> <b>Description: PIZZA</b> <b>Category: RISTORANTI</b>	<b>RISTORANTE LA FORCHETTA D'ORO</b> <b>BAR LA FORCHETTA D'ORO</b> <b>BAR RISTORANTE LA FORCHETTA D'ORO</b> <b>PIZZERIA LA FORCHETTA D'ORO</b> <b>LA FORCHETTA D'ORO</b>
---	--

Figure 1: Example of some fields in a record, and its formulation variants

A record, and the formulation variants generated, as described in [4], by a rule-based system for this record are shown in the top and bottom half of Figure 1 respectively.

Several assessment sessions were performed for evaluating the coverage of the user formulations by the FVs. In particular, a large database (DB20000) was collected selecting from a month and a half of customer calls to automatic systems operating in 14 call centers distributed in several regions of Italy. 20216 business calls, routed to the human operators by the system, because it failed to deliver the desired information, were selected and transcribed. The selected calls correspond to the most frequently asked listings. The phone number provided by a human operator was associated to the log of each call.

To generate new, more accurate, FVs, the transcribed denominations were analyzed, and new generation rules derived. The FVs that received most attention were those related to hospitals, social services, public utilities, communication and transportation agencies, and the like, because they account for the majority of the calls.

An analysis of the DA system failures has been done to discover the main causes of its errors. The errors can be grouped in three main classes:

1. User formulations are slightly different (due to articles, prepositions, etc.) with respect to the stored set of FVs.
2. User formulations differ because extra words or sentences are inserted, or due to the deletion of words, even though part of the information is still present.
3. User formulations are completely different with respect to the stored set.

We focus, in this work, on the errors of the first and third class. In particular, frequent errors of the third class for a specific entry can give an indication that the insertion of new formulations in the DA database is required for that entry.

### 4. Automatic learning of formulation variants

As introduced in Section 2, the recognition module of the system produces, together with the lexical constrained word hypotheses, the phonetic transcription of each utterance using a phone-looped model. The phonetic strings associated to a given phone number are, thus, automatic transcriptions of the user formulations for the corresponding business listing.

Table 1 shows, as an example, a small set of unconstrained phonetic transcriptions associated to the most requested phone number in the DB20000 database, corresponding to *FSInforma*, a widely used automatic call center for train timetable information, developed by CSELT, and managed by the Italian railways service provider Ferrovie dello Stato.

Table 1: Samples of transcriptions of user requests for the railway information service *FSInforma*

ufiCoinfoRmaZionilstaZiuneditaeni	
enomaladelataZaneditoRenopoRtanovelomRaveRda	
feRoviodelostato	
ifuoRmaZionifeRoiedelostato	
esaZionetiboRtina	fveRuilstato
skazione	oRaRiodetReni
tReno	fRovionalostato
nomaRomeRbefese	saZinoCentRale

These phonetic strings are widely different, and some of them can hardly be decoded. Recall, please, that these utterances were not completed by the automatic DA system for several reasons such as endpoint detection failures, extra-linguistic phenomena, restarts, low confidence scores, recognition errors due to the lack of a suitable transcription in the current database, etc. Another cause of system failures is that the user request was ambiguous, incomplete or embedded in a sentence, so that only several turns of dialog with the user allowed the operator to deliver the information. On the other hand, it is possible to detect in Table 1 phonetic sequences that are easily interpreted since they are correct or nearly correct transcriptions of a denomination such as <feRovio-delostato> and <skaZione> for “Ferrovie Dello Stato” and “Stazione” respectively, and several variants with relatively few phonetic distortions.

#### 4.1. Accuracy of the recogniser

The phone accuracy of the lexical unconstrained recognizer has been evaluated by aligning its results with the phonetic transcriptions of 1000 phonetically rich sentences randomly selected from the sessions included in the blocks 10 to 30 of the SPEECHDAT2 Italian database. These sentences include 7K words and 34K phonemes.

Table 2– Results of a phone-looped model recogniser

DB	Phone accuracy	Deletion rate	Insertion rate	Substitution rate
SPEECHDAT2	82.2 %	5.1 %	3.1 %	9.7 %
FIELD	67.8 %	11.7 %	8.2 %	12.2 %

The phone accuracy, and the deletion, insertion and substitution error rates are shown in the first row of Table 2. The recognizer uses a set of 27 – single state - stationary units modeling the stationary parts of the context independent phonemes (less affected by the phonetic context) and 348 – two states - transition units defining all the admissible transitions between the stationary units [2]. The grammar of the recognizer is a phone-looped model that forces a sequence of two stationary units to be connected by the corresponding transition unit.

In the second row of Table 2, the results are given for a field database that includes 27K words and 272K phonemes. The reduction of the phone accuracy can be explained because the SPEECHDAT2 database includes read speech and the phonetic transcriptions are accurate, while the field data include user requests for business listings that the automatic system was unable to satisfy. These utterances were manually transcribed not accounting for endpoint detection errors, extralinguistic phenomena, acronyms, etc.

The “word” accuracy of the recogniser is about 10% for this set of utterances: this is not surprising recalling the difficulty of the task, highlighted in the previous section.

The distance between two strings of phones is obtained by Viterbi alignment of the two strings using the log-probability of insertion, deletion and confusion among phones. These probabilities were trained using another set of field data, aligning each phonetic sequence with its corresponding correct transcription.

#### 4.2. Clustering and selection of new formulations

Our working hypothesis, confirmed by the experimental results was that, collecting a large number of requests for the same listing, there is high probability of obtaining clusters of phonetically similar strings, whose central elements, defined as the string that has the minimum sum of the distance from all the other elements of the cluster, are quite accurate phonetic transcriptions of (possibly new) user formulations.

For the most frequently requested phone numbers in the DB20000 database, each set of phonetic strings was clustered into similar subsets by using a furthest neighbour hierarchical cluster algorithm based on the mutual distance between each phonetic string. The set of phonetically similar utterances is detected by setting a threshold on the within-cluster distance. The clusters with few elements and large within cluster variance are discarded.

If a large enough database is available, it is possible to select significant clusters, characterized by high cardinality and small dispersion of the included phonetic strings.

For example, using the 458 formulations that were available for the phone number of service *FSInforma* in the DB20000 database, the procedure generated several phonetically similar clusters, only three of them, however, were significant according to a selection criterion related to the number of elements in the cluster (> 20 in this case) and to a low (< 4.0) dispersion of the elements within the cluster.

The central element of the three clusters is shown in Table 3. It is worth noting that, when the number of elements of a

Table 3 – Central elements of the three significant clusters related to the denomination *FSInforma*

Central element	System nearest variant	No of elements	Cluster variance	Distance
feRoviedelostato	feRoviedelostato	156	2.13	0.00
staZioneCentRale	staZionefeRoviaRia	198	3.27	3.22
staZione	staZionefeRoviaRia	25	1.9	4.43

cluster is large enough, the central element of the cluster gives a very good transcription of the required denomination. For the central elements in Table 3, good formulation variant candidates are the phonetic strings <staZioneCentRale> and <staZione> that are distant from the already present formulation <staZionefeRoviaRia>, while <feRoviedelostato> exactly matches a formulation already in the system

## 5. Assessment of the procedure

To assess the capabilities of our approach the dimensions of the DB20000 database are not sufficient, we need a much larger amount of phonetic strings.

### 5.1. Experiments with automated calls

Since it is, currently, cumbersome to associate the phonetic strings with the phone number that will be eventually delivered by the operator, rather than relying on the calls that the automatic DA failed to serve, in a first experiment described in [4], we clustered the phonetic strings of calls successfully processed by the automatic system. Using a database of 340K phonetic strings we assessed the quality of our clustering procedure, and its ability to produce, as a central element of a cluster, a formulation variant that has been included in the system.

### 5.2. Experiments with calls routed to the operator

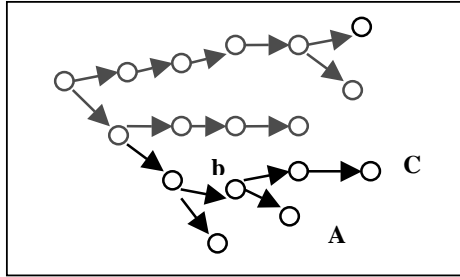
In the months of May and June 2001 a new database (DBMAY01) was collected including 7.2M phonetic strings related to business listing calls routed to the operators due to system failure to complete its transaction with the customer. The 7.2M calls in this database are distributed among the listings of 8100 different city names.

It is worth noting that the calls are routed to operators even if the recognized denomination is correctly found in the Top-List, but it has not a large enough confidence score, or if it could match the correct denomination in the complete set of listings, but the address is not known to the user.

Since for these calls we don’t have the phone number information associated with the phonetic string, but only city name, confirmed to the system by the user, we cluster all the strings associated with the same city name. The number of strings to be clustered can be as large as 723K for a big city like Rome.

For these experiments it was necessary to adapt our clustering procedure to deal with a logically huge, but very sparse distance matrix. The matrix is sparse because we are interested in clusters with small dispersion of the included elements, thus we ignore string distances greater than a small threshold. To compute the distance among a huge number of phonetic strings, we use a recursive tree to tree matching procedure, where a tree branch is a phonetic string, as sketched in Fig. 2. Since we impose the distance between two strings to be symmetric, a tree branch, representing the reference phonetic string, is matched only against the tree branches at its right using Viterbi beam search.

Fig. 2 - Tree of the phonetic strings



The trellis computed during the dynamic programming alignment of the reference branch **A** against the branches at its right is kept, up to the phone leading to node **b**, to minimize the computation load when matching branch **C** with the remaining branches. Although beam searching dramatically reduce the active nodes in the trellis, the computational load would still be high for very large trees like the one of big cities like Rome or Milan. Therefore, we partition the set of the phonetic strings of a city into subsets characterized by strings of similar length. In this experiment, strings of length  $L$ , are compared only with strings of maximum length  $L+4$ : this means that we allow warping paths with at most 4 insertions. This reduces the search space and the beam search work because it a priori excludes comparisons of strings with large length difference.

For the subset of phonetic string related to Rome, for example, we generated 50 trees, ignoring very long strings that would have very low probability of producing significant clusters.

Using a rather conservative approach for clustering, i.e. aggressive beam search pruning, we obtained about 27.1K clusters (and the phonetic string for their central elements) referring to listings of 2038 different cities. These clusters cover 525K calls, about 8% of the total calls. 5.2K of the detected clusters have a central element that is already covered by a system formulation variant in the TopList (its distance from the system FV is lower than a threshold of 1.0 that corresponds to slight phonetic variations).

To give an idea of the accuracy of the detected central elements, Fig. 3 shows the number of formulations (calls) that match a FV in the whole list of the FVs of a city within a phonetic distance.

It can be observed that 8523 detected FVs (62.0%) perfectly match a FV that is already included in the system. These formulations cover 234K (72.0%) requests.

The central elements covered by the whole set of system FVs is 13.7K. The remaining 13.4K central elements, covering 185K calls, have been filtered to eliminate few inappropriate formulations, such as “doctor”, or the generic term “hospital” for a big city.

After the filtering, we obtained a set of 12.3K new formulations covering 184K calls. Examples of new formulations derived from routed calls are reported in Table 4.

Thus, our procedure is able to filter 7.2M calls routed to the operators, and to detect a limited number of phonetic strings that can be inspected by human operators without experience in phonetics, because the strings are “readable”, i.e. can easily be orthographically transcribed and associated with the corresponding phone number. New formulation variants can be included in the system vocabulary or moved to the TopList.

Fig. 3: Number of formulations (calls) matching a system formulation variant within a phonetic distance

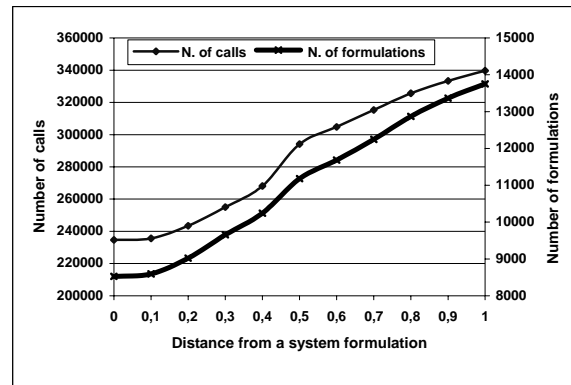


Table 4: New formulations derived from routed calls

Number of elements	Cluster central element	System formulation variant
511	ikea	dike
240	ufiCodiGene	azlufiCodiGene
121	CenRotumoRi	CentRoatoRi

## 6. Conclusions

We have shown that an unsupervised approach is able to detect user formulations that were not foreseen by the designers of a DA system. These formulations can be added to the system to reduce its failures. Conversely, the system formulations generated from the book listings for a given denomination, that never appeared in a huge amount of real data can be eliminated, to reduce the search costs, and substituted by frequently requested formulations.

Our procedures have been used to automatically detect new formulations, discover why existing formulations exhibit high rejection rates, and to compute the percentage of routed versus automated calls per phone number.<sup>1</sup>

## 7. References

- [1] R. Billi, F. Canavesio, C. Rullent, “Automation of Telecom Italia Directory Assistance Service: Field Trials results”, Proc. IVTTA 1998, Turin, pp. 11-16, 1998.
- [2] L. Fissore, P. Laface, F. Ravera, “Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition”, Proc. EUROSPEECH 95, Madrid, pp. 799-802, 1995.
- [3] V. Gupta, S. Robillard, C. Pelletier, “Automation of locality recognition in ADAS plus”, Speech Communication, vol. 31, n. 4, pp. 321-328, 2000.
- [4] C. Popovici, M. Andorno, P. Laface, L. Fissore, M. Nigra, C. Vair, “Directory Assistance: Learning User Formulations for Business Listings”, 2001 IEEE ASRU Workshop, Madonna di Campiglio, Italy, Dec. 2001.
- [5] H. Schramm, B. Rueber, A. Kellner, “Strategies for Name Recognition in Automatic Directory Assistance Systems”, Speech Communication, vol. 31, n. 4, pp. 329

<sup>1</sup> This work was partially supported by the EU ITS Project SMADA Speech Driven Multi-modal Automatic Directory Assistance